INTEGRATING DATA LAYERS TO SUPPORT *THE NATIONAL MAP* OF THE UNITED STATES

E. Lynn Usery, Michael P. Finn, and Michael Starbuck

U.S. Geological Survey
Mid-Continent Mapping Center
1400 Independence Road
Rolla, Missouri, USA 65401

usery@usgs.gov
mfinn@usgs.gov
mstarbuck@usgs.gov

Abstract

The integration of geographic data layers in multiple raster and vector formats, from many different organizations and at a variety of resolutions and scales, is a significant problem for *The National Map* of the United States being developed by the U.S. Geological Survey. Our research has examined data integration from a layer-based approach for five of *The National Map* data layers: digital orthoimages, elevation, land cover, hydrography, and transportation. An empirical approach has included visual assessment by a set of respondents with statistical analysis to establish the meaning of various types of integration. A separate theoretical approach with established hypotheses tested against actual data sets resulting in an automated procedure for integration of specific layers also has been implemented. The empirical analysis has established resolution bounds on meanings of integration with raster datasets and distance bounds for vector data. The theoretical approach has used a combination of theories on cartographic transformation and generalization, such as Topfer's Radical Law, and independent research concerning optimum viewing scales for digital images to establish a set of guiding principles for integrating data of different resolutions.

INTRODUCTION

The U.S. Geological Survey (USGS) has begun a new program for supporting the needs of the nation for topographic mapping in the 21st century. That program is referred to as *The National Map* and involves a vision of:

> information current, seamless national digital data coverage to avoid problems now caused by map boundaries, higher resolution and positional accuracy to better support user requirements, thorough data integration to improve the internal consistency of the data, and dramatically increased reliance on partnerships and commercially available data (USGS, 2002).

This vision includes the development and maintenance of eight data layers: transportation, hydrography, boundaries, structures, elevation, land cover, orthographic images, and geographic names. The data will be available over the World-Wide Web (WWW) and accessible for both direct viewing on the web and for download by users. Data will be comprised of the best available source and the USGS will depend on state, local, tribal, and other government and private industry to supply data. The USGS will become a data producer only in cases where no other data are available.

The problem of using data from such a variety of sources becomes one of integration of the various resolutions and accuracies of data in both horizontal and vertical directions. Data must be horizontally integrated to provide the seamless nation-wide coverage as specified and vertically integrated among the different themes to provide internal consistency. The data integration problem is one of massive proportion and the USGS has ongoing research to develop procedures to accomplish this integration (for example, see Finn *et al.,* 2004). It is the purpose of this paper to document some of our progress to date and to better define the exact nature of the data integration problems. Specifically, the next section will address the basic meaning of the term data integration in raster, vector, and combined geometric domains. The third section will detail our approach to this problem and the testing that has

occurred to determine limits of integration. The fourth section will document an approach for vector and raster integration based on transportation and orthographic images. A final section will conclude with our current ideas for a theory of integration based on the concept scale and resolution ratios, optimum viewing scales, and image fusion concepts.

DATA INTEGRATION DEFINITION AND VISUALIZATION OF THE PROBLEM

The concept of an integrated dataset of various layers is based on the approach used in the standard five-color lithographic topographic map, which the USGS has produced for decades and provided to its customer base. In the same way that all features of different types on the lithographic map are co-registered and integrated into a single document, digital data sets need to register and integrate in a similar fashion. A major difference is that the USGS produced all the data for the topographic map and could force resolution and accuracy limits to maintain an integrated product. In the current environment of *The National Map* data are provided by a variety of sources and at a variety of resolutions and accuracies. Forcing consistency is no small achievement and simply establishing the meaning of an integrated dataset poses difficulties. For example, Figure 1a shows transportation and an orthographic image in an area west of St. Louis, Missouri, USA. The image is a color orthophotograph with one-foot (0.33m) pixel size from Nunn-Lugar-Domenici 133 priority cities of the Homeland Security Infrastructure Program (Vernon, Jr., 2004), which approximates the resolution. The transportation file is from the Missouri Department of Transportation (MO-DOT) and provides one of the most accurate sources for this area. Note the mismatch between roads as shown on the image and roads from the vector data file. Is this an integrated dataset? We provide a second example in Figure 1b using the same area, same orthophotograph, but with Census Topologically Integrated Geographic Encoding and Referencing (TIGER) line files for a transportation source. The base source of the TIGER data is the USGS 1:100,000-scale topographic maps. As is evident in this example, the TIGER data is not integrated well with the image. Note that in both cases we really have not integrated the datasets, we have merely provided an overlay of the roads on the image. A final example is shown in Figure 2 including hydrography data overlaid on the same image base. In Figure 2a, the hydrography source is the USGS National Hydrography Dataset (NHD) while Figure 2b shows hydrography from St. Louis County. The St. Louis County data is certainly better and actually shows the streams as double lines, but it still doesn't match the image exactly. What does it mean to be integrated?

(a) (b)

Figure 1. MODOT transportation overlaid on an orthographic image is shown in (a) while Census TIGER transportation overlain on the same image is shown in (b).

We take the position that integration means the datasets match geometrically, topologically, and have a correspondence of attributes. Thus, from the vertical integration point of view, to be integrated, the vectors from the transportation and hydrography files in Figures 1-2 need to exactly follow the corresponding features in the images. Further, if we have such a match we can fuse the vectors into the image without loss of information since the vectors

will exactly align. From a horizontal integration viewpoint, two maps must share exact attribution so an extension of a feature from one horizontal partition to another, remains the same feature with the same attributes.

(a) (b)

Figure 2. Shown in (a) is hydrography from USGS NHD while (b) shows hydrography from St. Louis County.

Vertical and horizontal integration of vector and raster data are discussed above, but what does it mean to have two integrated raster datasets? For example, from *The National Map*, we use the USGS National Elevation Dataset (NED). This dataset includes data at 1, 1/3, and 1/9 arc-sec resolution (approximately 30, 10, and 3 m, respectively). The orthographic images for urban areas are one-foot resolution. If we integrate the elevation data, perhaps in the form of a shaded-relief presentation, with the image, we combine approximately 8100, 900, and 81, image pixels to match one elevation pixel (Figure 3). How do we know when two raster datasets are integrated? We can base it on the geometric frame of reference, but visually does it matter? In the case of a lake, the elevations should be flat and with flowing streams, the water should flow downhill, but can we really determine that with large resolution differences? One of the goals of our work has been to try to define the limitations based on resolution and accuracy, at which datasets can be realistically integrated.

Figure 3. An orthographic image with 0.33 m pixel size overlaid with elevation data with 30 m pixels.

We have a similar problem if we discuss integration of two vector datasets, but with *The National Map*, this is a simpler case since we only include vector representation of transportation, hydrography, and structure outlines. For transportation and hydrography, vertical integration should yield locations of bridges, culverts, and other structures. Resolution issues abound here as well, but accuracy appears to be a larger issue (Figure 4).

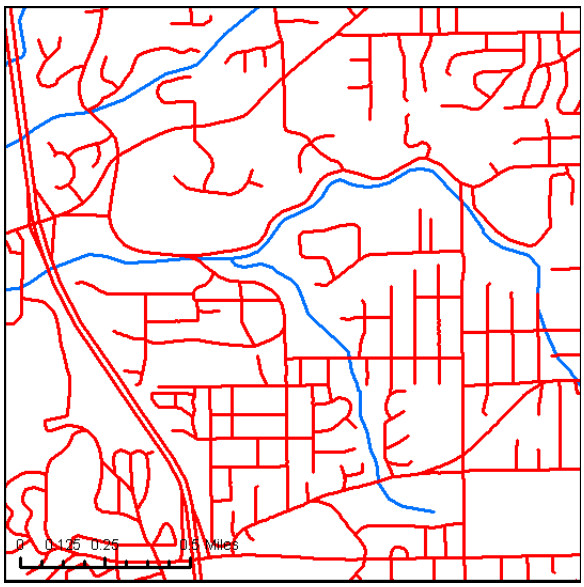

Figure 4. Vector data for roads (red) and streams (blue) overlaid for the same areas.

APPROACH AND STUDY AREAS

Our approach includes an empirical exploratory analysis to establish a meaning, both visually and numerically, for data integration; development of an hypothesis for data integration feasibility based on resolution and accuracy; and algorithmic development of procedures to shift features from one dataset to match a second to accomplish data integration. We selected five datasets and two test sites. The data include transportation, hydrography, land cover, elevation, and orthographic images (Table 1). We selected test sites over St. Louis, Missouri, and Atlanta, Georgia, USA based on the availability of the five data layers for testing.

Table 1
Study Data Layers

Data	Source	Type	Resolution	Accuracy	NM Layer
Elevation	NED	Raster	30 m	2-10 m	Elevation
Hydrography	NHD	Vector		13 m	Hydrography
Images	133 Urban Areas	Raster	1 ft	1 ft	Orthoimagery
Land Cover	NLCD	Raster	30 m		Land Cover
Transportation	Variable	Vector		varies	Transportation

The empirical testing was accomplished by overlaying one dataset on another, producing printed versions of the overlaid datasets, and conducting a visual analysis using a set of respondents to judge the effectiveness of the integration (match) between features in the two datasets. The hypothesis development is based on concepts from cartographic theory, including the Radical Law (Topfer and Pilliwizer, 1966), known limits of generalization methods, an empirical analysis of viewing scale (Fleming, *et al.,* 2005), and an examination of the results from image

fusion methods for remotely sensed images of varying resolution. The algorithmic development has followed the work of Chen *et al.* (2003a; 2003b) and attempts to force a vector transportation network to fit a corresponding image.

EMPIRICAL TESTING

For the five datasets in Table 1, we produced plots of all pairwise combinations at 1:24,000 and 1:12,000 scales. We asked a group of skilled cartographic technicians to judge whether the two datasets were integrated on a scale of 1 to 5, where 1 means no correspondence between the two datasets, 3 is moderate correspondence or integration, and 5 means perfectly integrated. The numbers 2 and 4 provided intermediate values in the scaling. These ratings were provided for three aspects of integration: position, shape, and temporality. Position is a measure of distance separating the same feature on the two sources. Shape assesses the correspondence of shapes but not necessarily alignment. Temporality is a judgment of whether the same feature exists on both sources. The respondents were shown examples of overlaid datasets meeting these ratings to help them gauge the correspondence. Table 2 presents a sampling of the results. The scores are a composite of the three measured aspects.

Table 2
Sample Results of Visual Interpretation of Integration

Paired Data Sources*	**12K Average Score (1-5)**	**24K Average Score (1-5)**
NLCD – NHD	1.2	1.1
NLCD – MO-DOT	1.0	Not evaluated
StierLC – NHD	1.2	1.2
Ortho – NHD	2.3	2.8
Ortho – TIGER	1.3	1.5
Ortho – MO-DOT	3.6	3.6
NED – NHD	1.0	1.3
NED – MO-DOT	1.0	1.0
StierLC – MO-DOT	3.0	3.4

*** NLCD – National Land Cover Dataset, NHD – National Hydrography Dataset, MO-DOT – Missouri Department of Transportation, StierLC – High resolution land cover, Ortho – Color Ortho image 1 ft.**

Our preliminary interpretations are that the results generally follow expectations regarding data resolution. Orthoimagery with a 1 foot resolution did well, especially when compared with MO-DOT vector transportation. The Ortho/TIGER results can be explained by the poor spatial registration due to the small scale source of the original TIGER data. The NED, with a 30m resolution, was hard to visually assess compared with the other data layers plus it is difficult to determine what to actually use to determine quality of feature registration. In general, the raster-to-raster overlays were not evaluated since there is no obvious basis for visual assessment.

AN APPROACH FOR VECTOR AND RASTER INTEGRATION

In trying to expand on the methodology for integrating vector data with orthoimagery, the USGS provided a grant to the University of Southern California (USC), Information Sciences Institute to fund, in part, continuing work on an automated road integration approach. USC's approach (Chen *et al.,* 2003a, 2003b) to a specific aspect of this problem of identifying road intersections from orthoimagery is as follows:

- Classify pixels as on-road/ off-road
- Compare to road network nodes (intersections)
- Filter algorithm to eliminate inaccurate pairs

This approach builds on the classic conflation techniques of Saalfeld (1993). The grant stipulations provided USGS researchers with the documented methods in this approach and the output files but not the source code.

USGS research scientists and computer programmers then designed and developed software to emulate USC's method for testing purposes and feasibility analysis of automated road integration. The fundamentals of the developed algorithm are as follows:

1 - Locate nodes (intersections) in vector data
2 - Create a buffer around nodes and create within each buffer an image template of road segments (geometrically accurate of attribute width)
3 - Drop the buffer template into the original raster imagery
4 - Perform pattern matching to identify the best match to the template
5 - Repeat steps 3 and 4 for all nodes in the vector data
6 - Filter out poorly identified intersections
7 - Perform rubber-sheeting transformation to correct the vector roads (for example, see Saalfeld, 1985)

Figure 5 shows an example of this localized template matching as described in steps 1 through 4 above. For testing purposes, we used the MO-DOT and the orthoimages from the 133 priority cities for the St. Louis area. We took assessments of the results of the automated road integration by both qualitative and quantitative methods.

Figure 5. Localized Template Matching (from Chen *et al.,* 2003b); the black dots show the vector nodes (road intersections) and the pixels converted to white show the best match to the template based on the pattern matching.

Quantitatively, the completeness and correctness were measured for the TIGER and MO-DOT data both prior to and after operating the algorithm on a portion of the St. Louis area. The algorithm increased the completeness from 25% to 64% for the TIGER data and from 93% to 98% on the MO-DOT, and increased the correctness from 27% to 69% for the TIGER data and 93% to 99% for the MO-DOT data (Chen *et al.,* 2003b). The lower the accuracy of the vector data, which is normally a function of source map scale, the greater the improvement the technique provided.

Qualitatively, we compared the output of the automated approach with the ideal result, which was constructed by manually editing the overlaid vectors to match the high-resolution orthoimagery. Figure 6 shows an enlarged portion of this ideal standard. Figure 7 shows a case where the USC algorithm improves the alignment for the road vector data. The visual assessment shows that this algorithm improved the alignment in most cases but, unfortunately, there were some cases where the algorithm caused degradation to the alignment.

Figure 6. Manually edited vector to match orthoimagery; the standard from which qualitative analysis of the automated road integration.

Figure 7. MO-DOT and orthoimagery integration with the USC algorithm showing improvement in alignment for

integration (red: MO-DOT; yellow: automatically processed roads).

Further qualitative analysis, focusing on the USGS algorithm, took into account the function of the type of control point filtering and the magnitude of the filtering. We looked at plots of a portion of the St. Louis area with 50% of the control points filtered using two different methods: a distance filter and a vector median filter. The distance filter eliminates control points identified by the algorithm solely on the difference of the distance, i.e., considering only magnitude, whereas, the vector median filter calculates a median vector of all control points and filters those points with the greatest difference between the control point and this vector, thus considering both direction and magnitude. A visual assessment of these plots appears to indicate that the vector median filter is preferable, but at this point this conclusion is tenuous. In addition, we compared the different percentages of points removed for the vector median filter method sequentially between 10% - 90% incrementing by 10%, and found that there is a more noticeable difference between 10% and 50% of points removed than between 50% and 90% of points removed.

We compared the USGS algorithm to the USC algorithms and found that, in general, the USC algorithm performed slightly better. We suspect that this is due to better image process/ pattern recognition and matching in the step that identifies road intersection on the imagery. For example, a visual assessment from the St. Louis area revealed that there was a road grid area in a section of the image where we observed that the USC algorithm appeared to be selecting better control points within the image. It appears that from a similar visual assessment from the Atlanta area, the USGS algorithm provides better results compared to the USC algorithm than it did in the St. Louis scene, though still not as good. This is likely attributable to the hazy nature of the Atlanta image.

This approach documented here illustrates a design for general vector/ raster integration based specifically on integrating vector road data with high-resolution orthoimagery. This design should support a variety of geospatial data and image sources.

CONCLUSION: TOWARDS A THEORY OF DATA INTEGRATION

Our project goal is to develop theory that can be used to implement an automatic method to support data integration based on available information about resolution and accuracy in metadata. This development is based on concepts from cartographic theory, known limits of generalization methods, an empirical analysis of viewing scale (Fleming, *et al.,* 2005), and an examination of the results from image fusion methods for remotely sensed images of varying resolution Our working hypothesis is that if scale denominators of source maps for vector data are within a factor of two, then the datasets can be integrated. If the factors are greater than two, then it may be possible to integrate the datasets, but significant processing and human interaction may be involved. For raster data, our working hypothesis is similar, but is based on a resolution ratio of two. This hypothesis is supported by work on viewing scales by Fleming *et al,* (2005), but is contravened by ongoing work by Ling and Usery (In review), that shows image fusion of satellite sources can be accomplished at resolution ratios of 1 to 30. Such large ratios do introduce artifacts, however, and the exact resolution ratio for true integration and image fusion without artifacts are in the process of being established.

REFERENCES

Chen, Ching-Chien, Craig A. Knoblock, Cyrus Shahabi, and S. Thakkar (2003a). Automatically and Accurately Conflating Satellite Imagery and Maps, In *Proceeding of the International Workshop on Next Generation Geospatial Information*. Cambridge, Massachusetts

Chen, Ching-Chien., Cyrus. Shahabi, and Craig . A. Knoblock (2003b). Automatically Conflating Road Vector Data with High Resolution Orthoimagey. *Report to U. S. Geological Survey on Grant No. 03CRSA0631*. University of Southern California, Los Angeles.

Finn, Michael P, E. Lynn Usery, Michael Starbuck, Bryan Weaver, and Gregory M. Jaromack (2004). Integration of *The National Map*. Abstract presented at the XXth Congress of the International Society of Photogrammetry and Remote Sensing, Istanbul, Turkey, July. Internet at: http://carto-research.er.usgs.gov/data_integration/pdf/integrationsISPRS.pdf. Accessed April 28, 2005.

Fleming, S., T. Jordan, M. Madden, E.L. Usery, and R. Welch (2005). "GIS Applications for Military Operations in Coastal Zones," In review, *ISPRS Journal of Photogrammetry and Remote Sensing.*

Ling , Y. and E.L. Usery (2005). “Assessment of Resolution Ratios of Input Images in Image Fusion,” Submission to *ISPRS Journal of Photogrammetry and Remote Sensing.*

Saalfeld, Alan. (1985). A Fast Rubber-Sheeting Transformation Using Simplical Coordinates. *The American Cartographer*, Vol 12, No. 2, pp. 169 – 173.

Saalfeld, Alan. (1993). *Conflation: Automated Map Compilation*. Computer Vision Laboratory, Center for Automation Research, University of Maryland.

Topfer, F. and W. Pilliwizer (1966). “The Principles of Map Selection,” *The Cartographic Journal*, 3, 10-16.

USGS (2002), *The National Map: Topographic Mapping for the 21st Century*, http://nationalmap.gov/nmreports.html, Accessed, April 25, 2005.

Vernon, Jr., Daniel E., (2004). Geospatial Technologies in Homeland Security. *EOM, Earth Observation Magazine*, Vol 13, No. 1 (Jan). GTIC America, Inc. Frederick, Maryland.